DOCUMENT RESUME

ED 177 211 TM 009 728

AUTHOR Winters, Lynn
TITLE Alternative Scoring Systems for Predicting Criterion Group Membership.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Apr 79
GRANT 08-NIE-G-78-0213
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Analytical Criticism; *Essays; Evaluation Criteria; *Evaluation Methods; *Expository Writing; Higher Education; Predictive Ability (Testing); *Scoring; Scoring Formulas; Senior High Schools; Student Ability; Test Reliability; *Writing Skills
IDENTIFIERS CSE Analytic Scale; Diederich Rating Scale; Holistic Evaluation; T Units; *Writing Evaluation

ABSTRACT

Four systems for scoring student essays were used to classify eleventh grade and undergraduate students according to writing ability. The reliabilities of the raters and the validities of the systems in classifying students were emphasized. Two analytic scoring systems--which assume that quality writing is characterized by the inclusion of certain rhetorical elements--were used: the Diederich Expository Scale and the Center for the Study of Evaluation (CSE) Analytic Scale. The T-unit analysis approach was used as an example of the frequency counting method; T-units were defined as the shortest grammatically complete units or sentences in a passage. The general impression, or holistic approach, was the fourth method used. Two samples of a 200-word expository essay were collected from students in four criterion groups: (1) non-college bound eleventh graders; (2) college bound students; (3) college students in a remedial composition course; and (4) college freshmen composition students. All of the systems--except T-unit--discriminated between the groups, particularly when three systems were used together. Classification accuracy was highest for the low-ability groups. Conclusions were made regarding the systems' generalizability and applicability. (Two writing topics are appended). (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED177211

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ALTERNATIVE SCORING SYSTEMS

FOR

PREDICTING CRITERION GROUP MEMBERSHIP

Lynn Winters
Center for the Study of Evaluation

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Lynn Winters

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 009 728

Paper Presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 8-12, 1979.

Session 8.24

The project reported herein was performed pursuant to a grant from the National Institute of Education (08-NIE-G-78-0213), Department of Health, Education and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

ALTERNATIVE SCORING SYSTEMS FOR PREDICTING CRITERION GROUP MEMBERSHIP

Lynn Winters

Center for the Study of Evaluation
UCLA

The method selected for judging writing samples *ipso facto* defines "good writing." However, decisions about writing competence derived from one scoring system may not lead to the same decisions about an examinee when another scoring system is used. The effect of using alternative methods of judging compositions on the classification of examinees into *a priori* defined levels of writing skill has not been investigated. The purpose of this study was to describe the effects of using competing scoring systems on classification decisions involving high school and college writers. Two questions were of interest:

1) What are the comparative reliabilities of different scoring systems?

2) What are the comparative validities of different systems for classifying writers into the appropriate levels of writing skill?

The Validation of Scoring Systems

Types of Scoring Systems

Scoring systems are generally of three types, impressionistic, analytic, or frequency counts (Braddock, et al., 1963). General impression marking is the system whereby two or more raters quickly read an essay then assign a single score ranking the writing sample in relation to all other papers being evaluated. The several ratings assigned by different scorers are averaged to obtain a reliable (i.e., stable) estimate of a writer's

ability. Several sorts of general impression marking are popular and incorporate rubrics of various levels of specificity and different score ranges. The one feature shared by impressionistic systems is the inclusion of rater training prior to scoring in order to "calibrate" readers in the assignments of marks (Conlan, 1976).

Analytic scoring systems, in contrast to general impression marking, require careful scrutiny of the writing samples. While impressionistic systems are predicated on the assumption that good writing is clearly recognizable, analytic systems assume that quality writing is characterized by the inclusion of certain rhetorical elements. One of the most popular of the analytic rubrics is the ETS Composition Scale, often called the Diederich Expository Scale. This scale emerged from a study that investigated factors influencing human judgment of expository writing samples. A factor analysis of patterns of rater agreement, combined with a review of rater comments on the essays belonging to each of five obtained clusters, was used to form the categories for the Diederich Expository Scale (DES).

The third type of scoring system is the frequency count. Whereas general impression and analytic scales focus on the quality of an essay, frequency counts are concerned with the quantity of such linguistic elements as the number of unique words in an essay, the number of sentences, or the number of words per independent clause. The frequency counting method with perhaps the most potential for judging levels of writing skill is the T-unit. T units are the shortest grammatically complete sentences that a passage can be cut into without creating fragments (Hunt, 1977). Studies investigating the relationship of T-unit length to some measure of writing quality have tended to support the hypothesis that a positive relationship

between the two exists (Howerton, 1977; Hunt, 1977; O'Hare, 1973).

A number of studies (Follman & Anderson, 1967; O'Hare, 1973; Howerton, Jackson, & Selden, 1977) have found relationships among the three types of scoring systems. These findings have led researchers to hypothesize that different scoring systems measure a number of elements in common (Follman & Anderson, 1967). However, a simultaneous comparision of impressionistic, analytic, and frequency count systems has not been done. The different assumptions about writing skill implicit in each type of rubric suggest that high intercorrelations may not exist. If such is the case, their interchangeability for judging writing competence is impugned.

Validation Procedures

There have been few attempt to validate direct measures of writing skill (i.e., sets of response criteria) beyond the level of content or descriptive validity. The content validity of the scoring criteria used in NAEP and the College Entrance Examination Board's Advanced Placement Examination was established by "expert judgment" (Lloyd-Jones, 1977). Neither of these well-established systems have been investigated as to their ability to predict success on subsequent tasks or classify examinees, issues in empirical validation.

While few guidelines exist in the literature on writing research for the empirical validation of scoring systems, techniques for establishing empirical validity have been described in the context of criterion referenced testing (Hambleton et al, 1975; Harris, 1974; Millman, 1974). Criterion referenced measurement technology offers two major strategies for investigating empirical validity: 1) The use of scores on established tests as a criterion to which newly-developed test items should predict and 2) The use of the performance data of criterion groups to validate the accuracy

of classification decisions to be made from the test.

The use of criterion groups to investigate the empirical validities of different scoring criteria provides an a priori classification system against which the criteria can be tested, i.e., , the ability to sort examinees representing different levels of writing skill into the correct writing groups. The criterion groups in this study were chosen to contain low and high performance high school and college writers. These four groups provided an operational definition of the construct "writing competence." The four scoring systems validated were selected to represent the three types of scoring criteria, impressionistic, analytic, and frequency counting. Two types of analytic scales, the Diederich Expository Scale and the CSE Analytic Scale, were tested. The Diederich Scale contains both impressionistic elements and elements tied to the requirements of exposition. The CSE Analytic Scale was derived from current composition textbooks and emphasizes the elements of writing taught in the schools. T-unit analysis was the method selected to exemplify the frequency counting approach. It was chosen because it has validated against qualitative measures of writing and was expected to have elements in common with impressionistic and analytic scales.

Procedures

Assignment of Papers to Raters

Two designs were used for assigning student papers to raters, one for scoring sample papers used to evaluate the effectiveness of rater training (Pilot Studies) and the second to examine the writing performance of the four criterion groups across the four scoring systems (Final Papers). Eighty subjects produced two expository essays on two test occasions, one week apart.

Two sets of four raters were trained in two scoring systems and graded essays after each training session.

The Pilot Studies were used to insure that all raters were assigning essay scores consistent with each other, to investigate topic effects and to obtain estimates of generalizability for several research designs for the final rating of papers. These studies were conducted four times, once after each rater training session. Twenty-four papers, chosen to represent all four criterion groups and both topics, were scored by all four raters. Two analyses of variance were run on the pilot scores, one to obtain intraclass coefficient alpha and the other to obtain estimates of variance components for subjects, topics, raters, and their interactions under a random model, fully crossed design. The coefficients calculated from these analyses are displayed in Table 1 below.

Table 1.

Alpha and Generalizability Coefficients: Pilot Data

Coefficient and Condition	Scoring System			
	GI	DES	CSE	T
Alpha: 4 Raters--Topics Fixed	.97	.82	.96	.98
Generalizability: 4 Raters, 2 Topics	.87	.88	.80	.65
Predicted G: 2 Raters, 2 Topics	.85	.79	.78	.47
Predicted G: 2 Raters, 1 Topic	.85	.87	.85	.68

Results of the pilot studies indicated that fairly high reliability could be obtained under a nested design with two raters reading one paper per examinee. The final rating design crossed subjects and topics with raters, but nested raters within topics. Each rater read only 80 papers, one from each subject, forty of which were Topic 1. and forty of Topic 2. Subjects and topics were randomly assigned to raters.

Figure 1.

Assignment of Writing Samples

Rater	Group			
	HSL n=20	HSH n=20	Coll L n=20	Coll H n=20
1	T_1 n=10	T_1 n=10	T_1 n=10	T_1 n=10
	T_2 n=10	T_2 n=10	T_2 n=10	T_2 n=10
2				
.				
.				
6				

Subjects

The subjects of the study were high school and college students enrolled in composition classes in June, 1978. Twenty (20) sets of writing samples were randomly selected from papers obtained from writing classes at a suburban, working class high school and a major urban university. Classes were chosen to represent two levels of writing skill, high and low, for both high school and college populations. The students whose samples

were selected became the criterion groups described below:

1. Low Performance High School Writers (HSL)

   Students enrolled in basic composition classes at the eleventh grade level who do not intend to attend a university.

2. High Performance High School Writers (HSH)

   Students enrolled in eleventh grade advanced composition classes who plan to attend a university.

3. Low Performance College Writers (Coll L)

   Students enrolled in remedial college composition at UCLA by virtue of having failed a written English placement examination.

4. High Performance College Writers (Coll H)

   Students enrolled in freshman composition at UCLA after having passed the written placement examination or a course in remedial composition.

Writing Tasks

Subjects were given 50 minutes to produce a 200 word expository sample on two test occasions, one week apart. Two parallel topics were randomly assigned to subjects and counterbalanced to control for test occasion (See Appendix). Dictionaries were not available, although subjects were told to revise papers if time allowed.

Controls

The design did not attempt to control for the instructional history of the subjects or the effects of test occasion. It did, however, attempt to control for the following: 1) topic effects 2) rater background 3) inter-rater disagreement. Table 2 and Figure 2 summarizes the steps taken to control sources of variation due to topics and raters.

Dependent Measures and Rater Training

There were four dependent measures for each subject. These consisted of each student's score over two topics and two raters for each of the four

scoring system, copies of the topics, and practice papers representing samples of each of the criterion groups. 2) Raters discussed the rubric and defined scoring categories among themselves. 3) Practice papers were scored and discussed until raters felt they could agree on the assignment of marks. 4) Raters read 24 sample papers for a pilot study (described in a previous section. 5) When the alpha coefficient from the pilot data was .80 or better, raters were deemed "trained" and allowed to score the actual writing samples.

In investigating the classification validities of the scoring systems, two issues were involved: Which systems discriminated the best among groups? And which systems classified subjects most accurately? For each of these issues, the dependent measures differed. The following outline summarizes the variables of interest for each question:

I. Variables of Interest in Finding the System(s) Which Best Discriminate Among Groups

   A. Independent Variables: Criterion Groups

   B. Dependent Variables: Scores Obtained on Systems

      1. GI: Score Scale 1-6
      2. DES: Score Scale 5-45
      3. CSE: Score Scale 6-24
      4. T: Score Scale 8-34 words per T unit

II. Variables of Interest in Finding Scoring Systems Which Best Predict Criterion Group Status

   A. Independent Variables: Scoring Systems as a Treatment

   B. Dependent Variables: Criterion Group Membership

Table 2.

A Summary of Controls for Sources of Variation Due to Topic and Rater

1. Examinee scores were averaged across topics

2. A stratified random sampling procedure was used so that each rater received on writing sample from each of the 80 subjects and an equal number of papers on each topic.

3. Papers were arranged in rater packages so that each rater received papers in random order. Additionally, rater packets were created so that there was no systematic pairing of raters.

4. Writing samples were typed and assigned random identification numbers to obscure examinee name and criterion group status.

5. To minimize reactivity, scoring systems were assigned to training conditions so that the two analytic scales were separated and general impression marking introduced first.

6. Each training condition consisted of four raters, two high school and two college teachers. All raters learned at least two systems. Two raters were trained in all four systems.

Figure 2.

Assignment of Raters to Training Conditions

	Session 1 GI & DES		Session 2 CSE & T
Rater 1:	College remedial teacher	Rater 1:	College remedial teacher
Rater 2:	College composition	Rater 2:	College composition
Rater 3:	High School English	Rater 5:	High School English
Rater 4:	High School English	Rater 6:	High School English

Analyses

Several preliminary analyses were necessary in order to select the appropriate procedures for analyzing data related to the question of classification validity. Rater agreement and generalizability coefficients were calculated (Appendix). The relationship among scoring systems was examined with Pearson Correlations (Appendix). Finally, univariate analyses of variance for each scoring system were run to identigy those systems which distinguished among writing groups (Appendix).

The results of these analyses indicated that while all systems had high alpha coefficients, the reliability coefficients were somewhat lower and not as comparable. T unit, which had the highest alpha (.99) could not be investigated with a G coefficient due to unusual subject by topic interactions. DES had the highest generalizability coefficient (.85) and an alpha of .90. GI and CSE were comparably reliable with G's of .63 and .67 respectively and alpha's of .81 and .97. Reliability dropped from the predicted pilot coefficients. Finally, unusual subject by topic effects of a high magnitude were found. These could not be explained because the research design confounded topic with occasion for the individual examinee and rater with topic.

Intercorrelations among the three systems for which significant differences among writing groups were found (GI, DES, CSE), at first appeared quite high. However, when the data were analyzed by criterion group, the relationship among these systems fluctuated from mild to weak. T unit analysis did not measure what is conventionally described by holistic systems (GI) or analytic criteria (DES, CSE) as good writing. Its negative relationship to the other systems for some of the criterion groups suggested

that, in the present study, information gained from T unit scores would be of little benefit in predicting expository writing performance as defined by criterion group membership. It was therefore not used in later analyses.

The classification validities of the systems were investigated by a simple ranking procedure and with discriminant analysis. The discriminant analysis (See Tables 3-4) revealed only one significant Linear Discriminant Function (LDF) for each of two combinations of scoring systems, GI/DES/CSE and DES/CSE. An examination of the within-group correlation matrices and the standardized LDF's indicated that DES possibly contributed most to group separation which it was combined with either CSE or GI. Both DES and CSE separated group centroids better than GI.

Table 3.

Within-Groups Correlation Matrix for Discriminant Analysis

System	GI	DES	CSE
GI	-		
DES	.54	-	
CSE	.48	.55	-
T	.01	.04	-.04

Table 4.

LDF Coefficients for Two Discriminant Analyses: GI/DES/CSE and DES/CSE

Analysis	Eigen Value	Relative % of Information Accounted for by LDF	Wilkes' Lambda	Standardized LDF Coeff.
GI/DES/CSE	3.388	96.69	.204	GI-0.061 DES-0.758 CSE-0.214
DES/CSE	3.374	99.69	.226	DES-.793 CSE-.232

Discriminant classification revealed a 61% accuracy of group placement when three systems were used (GI/DES/CSE) and a 59% accuracy with the two analytic systems. Classification accuracy was greatest for the High School Low group (95%) and least for the College High group (20%). Prediction was better for the Low writers in each level of school than the high. Misclassi ication tended to be from College High to College Low and College Low to High School High. The High School High group was misclassified as College HIgh when only the two systems were used. When three systems were used, misclassifications were almost equally distributed between the two college groups for the High School High Writers. The most accurate discriminant classification analysis is reproduced in Table 5 below:

Table 5.

Classification Accuracy Using GI, DES, CSE Scoring Systems

Actual Group	N	Predicted Group Membership HSL	HSH	Coll L	Coll H
High School Low	20	95%	5%	0%	0%
High School High	20	0%	55%	10%	35%
College Low	20	0%	25%	70%	5%
College High	20	10%	20%	45%	25%

Percent of cases correctly classified: 61.25%

The frequency distributions used for examining group classification for each scoring system separately displayed the same trends as the discriminant classification (See Figures 3-6). For all systems but T, the High School Low group was clearly discriminable from the other three groups. There was much intermingling among the other groups, although DES appeared

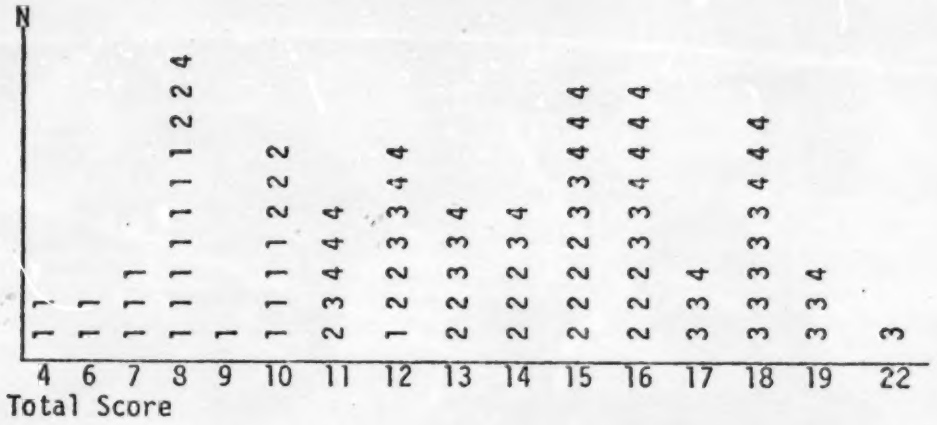

Figure 3. General Impression Scores Assigned to Criterion Groups

Group 1=HS L
2=HS H
3=Coll L
4=Coll H

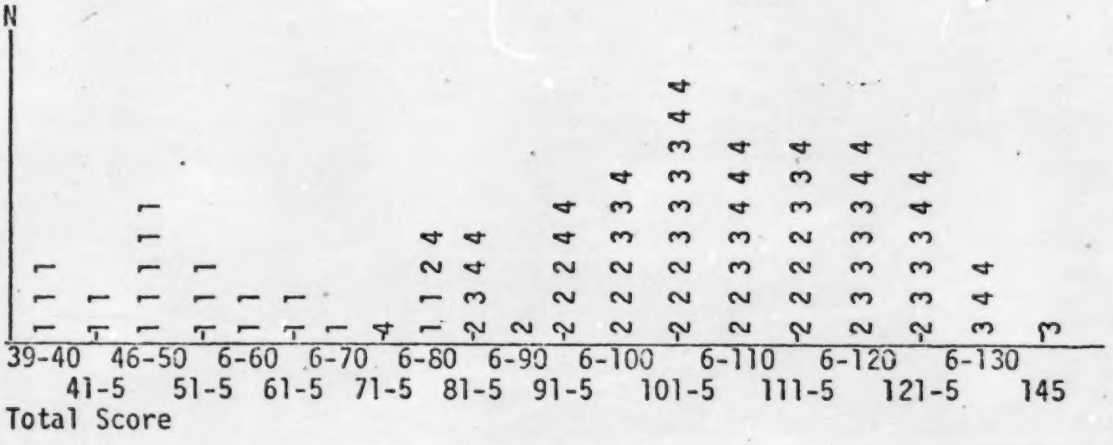

Figure 4. DES Scores Assigned to Criterion Groups

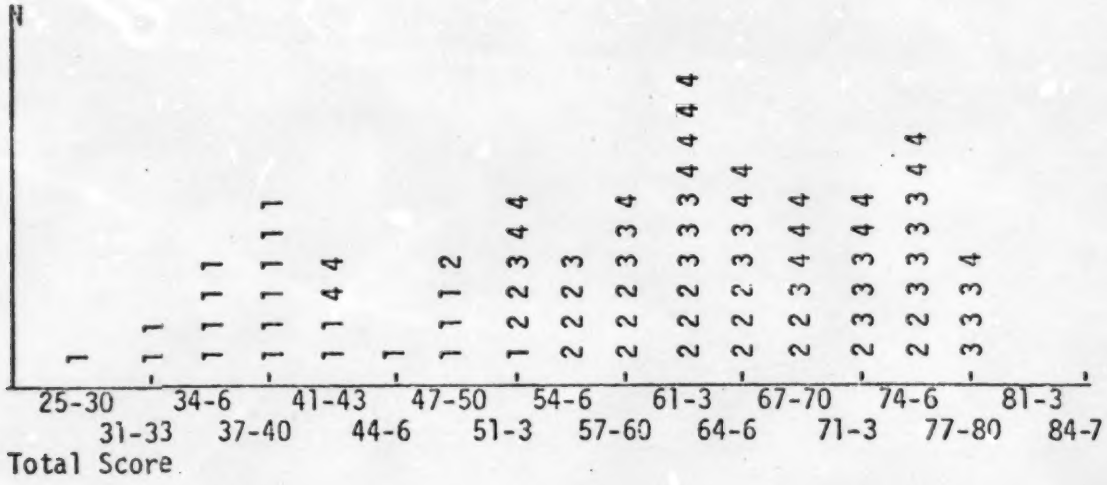

Figure 5. CSE Scores Assigned to Criterion Groups

Group 1=HS L
2=HS H
3=Coll L
4=Coll H

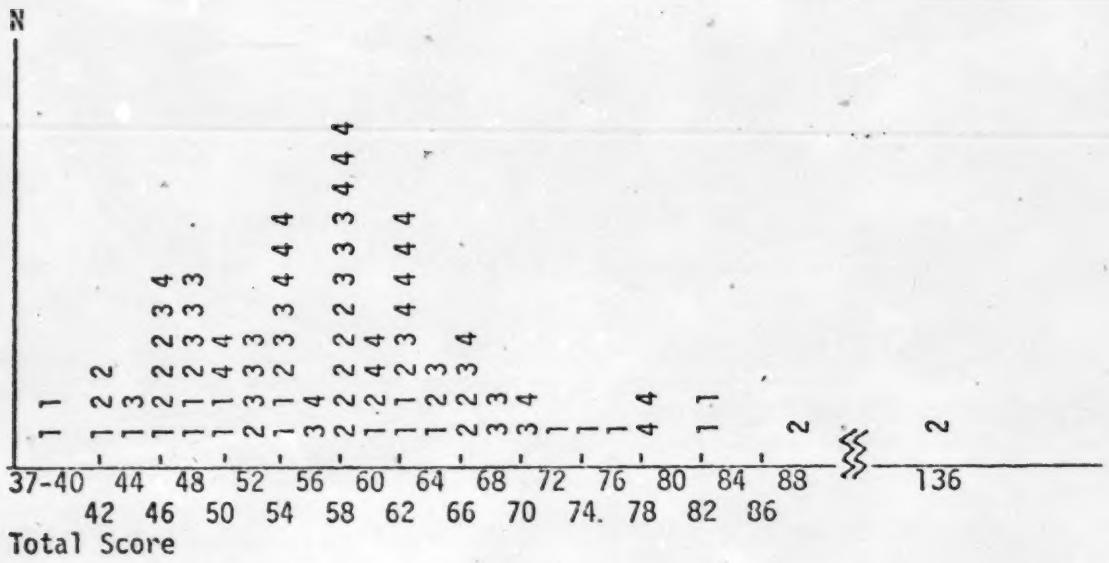

Figure 6. T-Unit Scores Assigned to Criterion Groups

Table 6

Classification Accuracy for Scoring Systems

Actual Group	n	Predicted Group Membership			
		HSL	HSH	Coll L	Coll H
(a) General Impression					
High School Low	20	81%	19%	0	0
High School High	20	15%	36%	46%	0
College Low	20	0	21%	29%	56%
College High	20	5%	28%	39%	36%
Overall GI Accuracy = 46%					
(b) Diederich Expository Scale					
High School Low	20	93%	8%	0	0
High School High		2%	46%	38%	16%
College Low		0	17%	35%	48%
College High		7%	3%	49%	37%
Overall DES Accuracy = 53%					
(c) CSE Analytic Scale					
High School Low	20	88%	13%	0	0
High School High		3%	46%	36%	51%
College Low		0	23%	36%	51%
College High		10%	19%	39%	33%
Overall CSE Accuracy = 51%					
(d) T-Unit Analysis					
High School Low	20	38%	18%	15%	30%
High School High		30%	17%	28%	25%
College Low		25%	34%	16%	25%
College High		8%	32%	41%	20%
Overall T-unit Accuracy = 24%					

to separate the High School High and College Low groups better than did the other two (GI/DES) effective scoring systems. With all systems but T, it appears the High School Low subjects are assigned the bottom third of the scores with the other criterion groups almost normally distributed about the top two-thirds of the score rankings and virtually indistinguishable from each other.

Unexpected findings include the impotence of T unit to discriminate or classify, as well as its non-existent relationship to qualitative scoring systems. Also unexpected was the wide divergence in types of writing samples receiving high and low ratings from each system. A perusal of the actual essays, more than any other analysis, illustrated the differences among the four systems. There was an unanticipated reversal in group means for the College Low and High groups. Finally, the inability of the systems, acting singly or in concert, to clearly separate all four groups was unanticipated.

Conclusions

Classification Validity

In order that a scoring system can accurately classify examinees into the correct writing groups, it must be able to distinguish among the groups. It was found that three of the systems, General Impression, the Diederich Expository Scale, and the CSE Analytic Scale were associated with differences in group performance. T unit was not, making it virtually useless for classification purposes.

When the three "discriminating" scoring systems (GI, DES, CSE) were used to classify examinees into the proper criterion groups, it was found that three systems classified more accurately than two. This finding was predictable in that classification accuracy is often enhanced by the use of several sources of information. Another explanation for the advantage of using three scoring systems to classify examinees is tied to the kinds of scoring systems used in this study. Two analytic and a general impression system were retained for the discriminant analyses. Although one analytic system (ES) contributed most to group separation, the best prediction was obtained with a comgination of DES and GI scores. It may be that accurate assessment of writing must incorporate both impressionistic and analytic strategies. Support for this interpretation is gained when it is remembered that the one system combining both general impression and analytic elements, The CSE scale does not lead to as accurate classification of group membership as DES. This result may be due, in part, to the fact that the scale provides no rating category for an overall impression of the effectiveness of the essay.

These interpretations must be accepted with caution, however, due to the unusual intercorrelations of the systems. There are groups by scoring system differences which may diminish the possibilities that any one scoring system can be an accurate predictor for all four groups. It is quite apparent that either holistic or analytic methods are more effective for distinguishing among writers of several ability levels and within a restricted age range.

The results of the classification analyses indicate that it is fairly easy to identify the lowest ability writers in a group and even, perhaps, some of the highest, but most difficult to separate writers of medium to high ability.

High school grouping practices provide a reason for the better classification accuracies obtained for the high school than college groups. In high school, students are assigned to English classes on the basis of previous grades, teacher recommendation, and post high school educational plans. The high school environment is smaller; teachers often know students well enough to "counsel" them into the appropriate English classes. On the other hand, the university composition placement decision is made on the basis of an impersonal, one-shot examination scored on criteria unknown to students and teachers. This procedure greatly increases the possibility for mistakes in classifying examinees.

Classification accuracy is best for the low writing groups in both high school and college. It may be that the characteristics of "low" writing performance are more stable and discernible than those of "high" writers. To paraphrase Tolstoy, poor essays are very much alike, but good essays are outstanding in their own ways.

Recommendations

The results of this study are constrained by several limitations on the generalizability of the findings. Scores reported for each system were obtained by raters who had participated in training. Rater training causes persons with potentially divergent views to "agree to agree" on both the meaning of the scoring rubric and the meaning of the scores. Different raters could agree with each other but disagree with the interpretation of the scoring rubric and score assignments reported in this study. The norming phenomenon, though clearly unavoidable, is rarely reported in writing research. The lesson: scoring rubrics never specify the entire set of criteria used to judge an essay. There is always that "x" factor which stands for how a rater interprets the rubric.

A second limitation on the generalizability of results has to do with the subjects. Participants in this study do not represent a random sample of high school or college writers. They do represent a sample of typical writing groups about which decisions must be made for placement into writing programs.

Finally, the topics used in this study, which appear to have produced some unusual effect on writing performance, were expository. There is no evidence that other modes of discourse or even other genres of expository topics would have produced the same results.

In spite of the aforesaid limitations, several recommendations, both methodological and substantive, emerged from the study. Several procedures used facilitated interpretation of results. The use of a generalizability study during pilot ratings made it possible to identify conditions responsible for inconsistencies in score assignment. The G study allowed various designs for the final study to be considered in terms of the trade off between

reliability and cost-effectiveness. The calculation of both rater and G coefficients clarified the amount of rater disagreement contributing to variation in examinee scores attributable to error. Rater training provided valuable insights as to how unman factors interact with scoring criteria and essays to produce a final estimate of writing ability. Finally, a review of representative essays revealed how scoring systems differed in their definitions of "good writing."

Substantive results point to testable generalizations about the comparative validities of the four systems for describing and classifying the writing performance of high school and college students. General Impression scoring, a widely used screening procedure which is speedily trained and scored, may not be the best system for making placement decisions or for graduation competency. The lack of descriptive power of its rubric severely restricts the instructional utility of this system when compared to analytic scoring techniques. Results of this study indicated that, while not sufficient as a placement tool, GI scores may be a necessary part of any writing assessment procedure.

The Diederich Expository Scale, while requiring extensive training time, appears to be the best system for distinguishing between high and low performance writers at both the high school and college levels. Its mild correlations with GI and CSE analytic scores indicate that the DES may be providing two sorts of information, impressionistic as well as analytic. It was found that discrimination and classification accuracy were increased when the number of data sources were increased. These findings suggest that the complex construct "writing skill" might best be assessed by more than one approach.

The use of the CSE scale to distinguish among examinees at different writing skill levels is a highly reliable and promising technique. The fact that this scale contributes less to group discrimination than the DES is possibly explained by its lack of attention to any of the qualitative aspects of writing. A revision of the scale with a category for the holistic aspects measured by GI may increase the classification and descriptive powers of this system.

The ability of measures of syntactic maturity is assess writing quality is seriously questioned by the results of this study. The lack of correlation of T-unit scores with three qualitative systems indicates that claims for a relationship between writing quality and syntactic maturity, at least for writers within a restricted age range, must be seriously re-examined.

References

Braddock, R., Lloyd-Jones, R., and Schoer, I. Research in written composition. Urbana, Illinois: National Council of Teachers of English, 1963.

Conlan, G. How the essay in the CEEB English test is scored. Princeton, New Jersey: Educational Testing Service, 1976.

Follman, J.C. & Anderson, J.A. An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English, 1967, 190-200.

Hambleton, R., Swaminathan, H., Algina, J., & Coulson, D. Criterion referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1977, 48, (1).

Harris, C.W. Some technical characteristics of mastery tests. In C.W. Harris, M.C. Alkin, & W.J. Popham (eds.) Problems in criterion-referenced testing. CSE Monograph Series in Evaluation, No. 3 Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Howerton, M.C., Jacobson, M. & Eldon, R. The relationship between quantitative and qualitative measures of writing skill. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April, 1977.

Hunt, K.W. Early blooming and late blooming syntactic structures. In C. Cooper and L. Odell (eds.), Evaluating Writing. State University of New York at Buffalo, 1977.

Millman, J. Criterion-referenced measurement. In W.J. Popham (ed.) Evaluation in education: Current applications. Berkeley, California: McCuchan, 1974.

O'Hare, F. Sentence combining: Improving student writing without formal grammar instruction. NCTE Research Report No. 15. Urbana, Illinois: National Council of Teachers of English, 1973.

APPENDICES

1. Topics
2. Data for Estimating Generalizability Coefficients: Final Papers
3. Intercorrelations Among Scoring Systems
4. Criterion Group Performance Across Scoring Systems: Means and Standard Deviations
5. Analyses of Variance for Each Scoring System

TOPIC 1

Reading, writing, and arithmetic, the three R's, have become a serious problem for many young adults. In fact, every year thousands of American students graduate from high school unable to read simple words, write short essays, or do math. To understand this problem, experts have studied areas such as teaching methods, family life, and the effects of television. However, no one best explanation has been found.

Assume you have been asked by a group of parents and teachers to explain the problem. Write an essay in which you explain why many of today's students graduate from high school without basic skills. Your essay should focus on one (or more) of the "three R's", reading, writing, or math.

Remember, the purpose of your essay is to give an informative explanation, not opinions. Back up your ideas with specific support, such as examples, facts, and other details.

TOPIC 2

The use of alcohol, drugs, and cigarettes, has become a serious problem for many young adults. In fact, every year the use of these substances hurts the lives of thousands of American teenagers by making them emotionally and physically ill. To better understand this problem, experts have studied areas such as family and school life, pressure from friends, and the effects of television and advertising. However, no one best explanation has been found.

Assume you have been asked by a group of parents and teachers to explain the problem. Write an essay to help explain why many of today's teenagers abuse alcohol and drugs or smoke cigarettes. Your essay should focus on one (or more) of the substances, alcohol, drugs, or cigarettes.

Remember, the purpose of your essay is to give an informative explanation, not opinions. Back up your ideas with specific support, such as examples, facts, or other details.

Table A-1

Data for Estimating Generalizability Coefficients: Final Papers

Source	SS	DF	MS	Estimated Variance Components*	
	(a) General Impression				
Subjects	224.286	57	3.935	.418	
Topics	3.446	4	.862	0.000	
Raters within Topics	3.941	8	.492	.003	G=.63
Subjects X Topics	55.304	57	.970	.325	
Residual Subjects by Rates within Topics	36.556	114	.321	.321	
	(b) Diederich Expository Scale				
Subjects	9649.122	52	185.560	33.473	
Topics	647.993	4	161.998	0.000	
Raters within Topics	76.227	8	9.528	.073	G=.85
Subjects X Topics	1202.453	52	23.124	8.852	
Residual Subjects by Rates within Topics	563.693	104	5.420	5.420	
	(c) CSE Analytic Scale				
Subjects	2447.302	62	39.473	4.400	
Topics	30.800	4	7.7	0.000	
Raters within Topics	23.092	8	2.887	.025	G=.67
Subjects X Topics	545.944	62	8.806	3.773	
Residual Subjects by Rates within Topics	156.376	124	1.261	1.261	
	(d) T-Unit Analysis				
Subjects	1370.150	51	26.866	0.000	
Topics	290.958	4	72.740	.002	
Raters within Topics	2.042	8	.255	.001	G= Undefined
Subjects X Topics	1488.059	51	29.178	14.493	
Residual Subjects by Rates within Topics	19.537	102	.192	.192	

*Negative variance components are reported as zero.

Table A-2

Intercorrelations Among Scoring Systems
for Total Sample (N=80)

System	GI	DES	CSE	T-Units
GI	-			
DES	.82*	-		
CSE	.79*	.86*	-	
T-Unit	.00	.06	.00	-

*Indicates significant correlation at $p<.05$, df 18.

Table A-3

Intercorrelations Among Scoring Systems
by Criterion Groups (n=20)

	Low			High		
	GI	DES	CSE	GI	DES	CSE
High School						
GI						
DES	.64*			.54*		
CSE	.71*	.79*		.44*	.30	
T	.28	.33	.30	.00	.24	.21
College						
GI						
DES	.53*			.51*		
CSE	.11	.37		.70*	.60*	
T	-.07	-.39	-.52	-.31	-.23	-.52

*Indicates significant correlation at $p<.05$, df 18.

Table A 4

Criterion Group Performance Across Scoring Systems:
Means and Standard Deviations

System	Group HS Low n=20	HS High n=20	Coll Low n=20	Coll High n=20	Total n=80
GI (Total 6)	$\bar{X}$ = 1.97	3.21	4.01	3.33	3.02
	S.D.= .50	.65	.72	.74	1.00
DES (Total 45)	$\bar{X}$ =13.25	25.66	28.00	26.00	23.29
	S.D.= 2.95	2.91	3.26	4.09	6.63
CSE (Total 24)	$\bar{X}$ = 9.84	15.50	17.07	16.18	14.65
	S.D.= 1.57	1.92	2.08	2.83	3.55
T-unit (Range 10-34)	$\bar{X}$ =14.18	15.07	13.99	14.89	14.52
	S.D.= 3.45	5.18	1.95	2.05	3.38

Table A5

Analyses of Variance for Each Scoring System:
Criterion Groups as Independent Variables

System	Source	SS	DF	MS	F	Eta^2
GI	Between	752.637	3	250.879	36.094	.59
	Within	528.250	76	6.951		
DES	Between	42930.338	3	14310.113	80.522	.76
	Within	13506.550	76	177.218		
CSE	Between	10280.700	3	3426.090	46.320	.65
	Within	5622.500	76	73.980		
T	Between	263.740	3	87.91	.470	.02
	Within	14205.085	76	186.90		